20

Towards Solving the Mixing Problem in the Decomposition of Geophysical Time Series by Independent Component Analysis

Filipe Aires

Department of Appl:ed Physics, Columbia University, NASA Goddard Institute for Space Studies, New-York, USA

William B. Rossow

NASA Goddard Institute for Space Studies, New-York, USA

Alain Chédin

CNRS Laboratoire de Météorologie Dynamique, Palaiseau, France

Short title: DECOMPOSITION OF GEOPHYSICAL TIME SERIES

Abstract. The use of the Principal Component Analysis technique for the analysis of geophysical time series has been questioned in particular for its tendency to extract components that mix several physical phenomena even when the signal is just their linear sum. We demonstrate with a data simulation experiment that the Independent Component Analysis, a recently developed technique, is able to solve this problem. This new technique requires the statistical independence of components, a stronger constraint, that uses higher-order statistics, instead of the classical decorrelation, a weaker constraint, that uses only second-order statistics. Furthermore, ICA does not require additional *a priori* information such as the localization constraint used in Rotational Techniques.

1. Introduction

This work concerns methods for the identification of the physical causes of variability of a given dynamical system from observations of its behavior. In many cases, an observed time series is produced by a mixture, linear or nonlinear, of different components representing different physical phenomena. In the linear case, the time series $\boldsymbol{x}(j)$ with temporal dimension, N, at particular spatial coordinate (we will call this a pixel), $j \in \{1, \ldots, M\}$, where M is the spatial dimension, is decomposed in time as:

$$\boldsymbol{x}(j) = \boldsymbol{G} \cdot \boldsymbol{\sigma} = \boldsymbol{g}_1 \sigma_1(j) + \boldsymbol{g}_2 \sigma_2(j) + \ldots + \boldsymbol{g}_Q \sigma_Q(j), \tag{1}$$

where the temporal base functions, $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_Q$, the columns of matrix $\boldsymbol{G}$, are unknown time series describing a fixed dynamical behavior (vectors and matrices are indicated in bold characters). In this paper, we consider decomposition in time, but the following discussion would be the same for a decomposition in space. Each $\boldsymbol{g}_i$ could be a signal with a different physical cause operative in a particular geographical region represented by a different score map $\{\sigma_i(j)\ ;\ \ j = 1, \ldots M\}$. One goal of an analysis is then to infer the unknown contributing components from the observed data, $\boldsymbol{x}$. In the linear case

$$\boldsymbol{h} = \boldsymbol{J} \cdot \boldsymbol{x} \simeq \boldsymbol{\sigma}, \tag{2}$$

where $\boldsymbol{J}$ is an estimate of the unknown matrix $\boldsymbol{G}^{-1}$ and $\boldsymbol{h}$ is an estimate of the unknown vector $\boldsymbol{\sigma}$. Statistical analysis methods that estimate $\boldsymbol{J}$ and $\boldsymbol{h}$ are called component extraction techniques. Their ability to retrieve good estimates, $\boldsymbol{h}$, of the true components, $\boldsymbol{\sigma}$, is highly dependent on the quality of the statistical dataset used (i.e.

sufficiently large number of independent examples sampling all the variations involved) and on the technical assumptions that are made about $\boldsymbol{J}$ and $\boldsymbol{h}$.

A common approach is to require the decorrelation of the extracted components (i.e., any two components are orthogonal): the covariance matrix of extracted components $< \boldsymbol{h}^t \cdot \boldsymbol{h} >$ is constrained to be diagonal; but this decorrelation principle has an infinity of solutions:

$$\boldsymbol{J} = \Theta \cdot \boldsymbol{J}_0, \tag{3}$$

where Θ is an undetermined $Q \times Q$ matrix so that $\Theta^t \cdot \Theta = \boldsymbol{I}_{Q\times Q}$. $\boldsymbol{J}_0 = \Sigma^{-1/2} \cdot \boldsymbol{E}^t$ is a $Q \times N$ matrix with Σ the truncated diagonal matrix of the higher eigenvalues of $< \boldsymbol{x}^t \cdot \boldsymbol{x} >$ and $\boldsymbol{E}$ the $N \times Q$ matrix with the associated normalized eigenvectors in the columns.

One particular decorrelation solution is the well-known Principal Component Analysis (PCA) or, in the geophysical community, Empirical Orthogonal Function (EOF), first used in atmospheric sciences by Lorenz (1951). In this technique, an additional constraint is added to resolve the indeterminacy of the decorrelation solutions: the successive extracted components have to explain the maximum remaining variance. This solution is given by taking $\Theta = \boldsymbol{I}_{Q\times Q}$ in Equation (3). Three problems could arise in using the PCA teechnique. (1) Even if the mixing of the components is linear as in Equation (1), this maximum-explained-variance assumption can cause mixing of different physical phenomena in the extracted components (Kim and Wu, 1999) as we will show here (see Figure 1-A for an schematic illustration of this problem in a

2-dimentional case). (2) This mixing problem is also particularly serious when the PCA is applied to data that have more than one component with about the same variance. In this case, the problem is not solvable since any orthogonal rotation of the principal components will be a PCA solution (Figure 1-B). (3) Since PCA imposes the orthogonality of the base functions it extracts, mixing problems also arise when the actual physical base functions are not orthogonal (Figure 1-C).

The PCA assumptions (linearity, variances explained by the components or orthogonality) used to resolve the solution indeterminacy are not known, a priori, to be valid for a particular dataset. It these PCA assumptions are not valid, variations that are not physically connected could be artificially gathered together into one extracted component. This is the reason why PCA is often used in restricted geographical domains instead of global domains to try to isolate a single dominant mode of variation, wich PCA can correctly identify. Consequently, even if PCA is useful as a tool for compressing information by **describing** the most variance with the fewest terms in an expansion, it can lead to misinterpretation of physical relationships.

Rotational Techniques (RT) were introduced (Horel 1981, Richman 1981), in part, to obtain a solution more physically interpretable. In these approaches, an additional constraint of localization, based on the so called "simple structure" principle, is used to solve the indeterminacy of the decorrelation solutions. There exist many proposed criteria for this purpose: quartimax, varimax, transvarimax, quartimin, oblimax, *etc* (Richman 1986). Two distinct classes of RT solution could be distinguished: the Orthogonal Rotations that preserve the orthogonality constraint of the components, and the Oblique Rotations that relaxe this constraint. However, the localization criteria used in both family of solutions are quite subjective and use of such *a priori* information may not be well suited to all applications.

The Independent Component Analysis (ICA) method, briefly described in Section 2, is based on information theory and has been recently developed in the context of signal processing studies and of the development of neural coding models (Bell and Sejnowski 1995). The two major distinctions between the ICA approach and the classical techniques are:

- The assumption of a linear mixture model is not required so an orthogonality constraint is not applied. However, the example presented here happens to be one where the signal is composed of the linear sum of components as in (1).
- The method extracts statistically independent components, even if these components have a non-Gaussian probability distribution function, by making use of higher order statistics, whereas the PCA or RT approaches use only second-order statistics.

We argue that the ICA approach is a particularly promising technique which may overcome the main pitfalls of the standard techniques (PCA or RT) for geophysical time series analysis. When faced with observations of a system with unknown dynamics, identifying the statistically independent variation mode seems more likely to be meaningful than assuming, a priori, that the system is composed of linearly mixed, orthogonal modes unless such modes ca be shown to be present form other information (North 1984). Moreover, if one or two modes are not known, a priori, to be dominant, then maximizing the variance explained by each mode as in PCA produces inappropriate mixing. Using other subjective criteria also seems difficult to justify. Hence, the classical techniques can be used in situations where additional information is available to justify their assumptions as useful approximations, but they cannot be used to search for new understanding of the system dynamics because they make such strong assumptions about it. ICA, by finding statistically independent modes, may provide a better way to explore the dynamics of a system, like the atmosphere-ocean system, that is known to involve non-linear coupling of many modes across a wide range of space-time scales; however, even statistical independence is only a guide to the system's behavior.

A previous study that applies ICA to the analysis of variations of tropical sea surface temperature (Aires et al. 2000) illustrates the potential of the ICA technique to separate a geophysical time series into more meaningful components. To illustrate more clearly how ICA handles the component mixing problem, we construct a synthetic dataset, where the true answer to the decomposition problem is known, and apply both PCA and ICA to extract components (Section 3). We deliberately devise a dataset to test whatever PCA can separate distinct modes of variation that are added linearly as many practitioners appear to expect. We show that, even in the case of a linear sum of components, the PCA technique mixes the contributions, but that the ICA method can correctly separate the components without additional subjective constraints like those used in RT.

2. The Independent Component Analysis technique

In this section, we briefly review the main concepts underlying the Independent Component Analysis (ICA) technique. For more details, the interested reader is referred to Bell et al. (1995) and Aires et al. (2000). The ICA technique aims to extract statistically independent components, a stronger constraint than the decorrelation requirement of the classical techniques. The statistical independence of two variables h_1 and h_2 is determined when their joint distribution can be factored:

$$P(h_1, h_2) = P(h_1) \cdot P(h_2). \tag{4}$$

This constraint involves higher-order statistics whereas the decorrelation constraint only involves second-order statistics. Decorrelation is equivalent to statistical independence only in the Gaussian case. So the higher-order statistics are particularly important when the analyzed data have components with non-Gaussian distributions. Avoiding the *a priori* assumption that second-order statistics are sufficient is important when the components are unknown as is usually the case. It is also important to not confuse the non-Gaussian character of the components with the non-Gaussian character of the data itself; however, if the data have a non-Gaussian distribution, then at least one component is also non-Gaussian, since for the simplest linear mixture of Gaussian components, the distribution would be Gaussian (a non-linear combination of Gaussian distributions could be non-Gaussian). Some previous studies examine this non-Gaussian behavior in the data (Burgers and Stephenson 1999, Aires et al. 2000). Without *a priori* information on this matter, the use of ICA is recommended since its requirement of statistical independence is more general than the decorrelation assumption.

The time series observations are gathered into a dataset $\boldsymbol{X}_j{}^t$ of M observations $\boldsymbol{x}(j) = (x_j{}^t \ ; \ \ t = 1, \ldots, N)$ with $j \in \{1, \ldots, M\}$, where M is the spatial dimension of the time series and N is its temporal dimension. The times series $\boldsymbol{x}(j)$ is assumed to be a mixture, linear or nonlinear, of several statistically independent components $\boldsymbol{\sigma} = \{\sigma_i \ ; \ \ i = 1, \ldots, Q\}$:

$$\boldsymbol{x}(j) = \mathcal{A}(\boldsymbol{\sigma}(j)) \tag{5}$$

where $\mathcal{A}$ is an unknown mixture function, which is, by hypothesis, non-singular (i.e. it can be inverted).

The goal of ICA is to retrieve a function $\Phi : \boldsymbol{x} \rightarrow \boldsymbol{h}$, where $\boldsymbol{h}$ is an estimate of $\boldsymbol{\sigma}$ and the terms $\{h_i \ ; \ \ i = 1, \ldots, Q\}$ are statistically independent. The estimate, $\boldsymbol{h}$, is defined as a deterministic function (linear or not) of the observations:

$$h_i = \Phi_i(\boldsymbol{W}_i, \boldsymbol{x}) \ ; \ \ i = 1, \ldots, Q \tag{6}$$

where $\{\boldsymbol{W}_i \ ; \ \ i = 1, \ldots, Q\}$ is the set of parameters of Φ. The number of components, Q, is here supposed to be known (this number can be estimated by a break in the frequency spectrum of the data, for example). With real observations, Q depends on the analysis objectives: extracting a lot of components allows for more complete description of the variability but makes the interpretation much more complicated, whereas extracting fewer components focuses attention on fewer different phenomena at the cost of explaining less of the variability. The interested reader should refer to an article by Nadal et al. (1999).

The parameters, $\boldsymbol{W}_\iota$, are estimated by applying a gradient descent algorithm to a cost function that specifies the statistical independence of the $\{h_i \ ; \ \ i = 1, \ldots, Q\}$. Different equivalent cost functions could be used; we focus here on the *infomax* approach to ICA (Nadal and Parga 1994) from which simple algorithms have been derived (Bell and Sejnowski 1995). Information theory is used to specify the statistical independence cost function: the fundamental quantity used here is *information redundancy*. Given Q variables, $h_1, h_2, \ldots, h_Q$, the information redundancy $\mathcal{R}(h_1, h_2, \ldots, h_Q)$ is defined as the Kullback diverg(nce between the joint distribution $P(h_1, h_2, \ldots, h_Q)$ and the factorized distribution $P(h_\iota) \cdot P(h_2) \ldots P(h_Q)$:

$$\mathcal{R}(\boldsymbol{h}) = \int_{-\infty}^{+\infty} \prod_{i=1}^{Q} dh_i \; P_h(\boldsymbol{h}) \; \log \frac{P_h(\boldsymbol{h})}{\prod_{i=1}^{Q} P_i(h_i)} \qquad (7)$$

This informatioıı redundancy comes from information theory and measures the difference between the joint and the factorized distribution: when the redundancy $\mathcal{R}(\boldsymbol{h}) = 0$, $P_h(\boldsymbol{h}) = \prod_{i=1}^{Q} P_i(h_i)$, which means, by the definition in equation (4), that the components of vector $\boldsymbol{h}$ are statistically independent.

The use of a gradient descent algorithm to minimize this cost function is interesting since it allows for the introduction of any *a priori* information about the solution that may be available. For such a purpose, a second term in the quality criterion is added that represents any additional constraint(s) on the solution. For example, this additional information could be a constraint on the shape of the solutions, on the distance of the solution from a first guess, or on the regularity properties of the solution. Such a regularization approach, also used in variational assimilation methods for example, is a classical way of using all the information that is available.

A nonlinear regression model for the extraction model in Equation (6) has to be specified. The Multi-Layer Perceptron (MLP) is often chosen: this artificial Neural Network model is preferred for its nonlinear behavior. In this experiment, we use a simple MLP architecture with no hidden layer. This neural mapping is defined by (from right to left in Figure 2):

$$\boldsymbol{y} = f(\boldsymbol{h}) = f(\boldsymbol{J} \cdot \boldsymbol{x}), \tag{8}$$

where f, the logistic function, is only used for algorithmic considerations. The extracted components are not the output $\boldsymbol{y}$ or the neural mapping (8), but the vector $\boldsymbol{h} = \boldsymbol{J} \cdot \boldsymbol{x}$. We use this model because the mixture model is linear (the nonlinear mixture case will be the subject of future work).

With the redundancy reduction criterion and no-hidden-layer architecture, an algorithmic implementation of the ICA has been found (Bell and Sejnowski 1995):

$$\Delta J_{ik} \propto J_{ik} + \bar{y}_i \cdot \sum_l J_{lk} \cdot h_l \tag{9}$$

where

$$\bar{y}_i = \frac{\partial}{\partial y_i} \frac{\partial y_i}{\partial h_i} = \frac{\partial}{\partial h_i} ln(\frac{\partial y_i}{\partial h_i}) = 1 - 2y_i. \tag{10}$$

This algorithm is described in a more practical way in the appendix. [1]

[1] See also the web page http://www.cnl.salk.edu/ tewon/ica_cnl.html of the Computational Neuroscience Laborabory of Terry Sejnowski at The Salk Institute for links to recent literature, software and demos concerning the ICA paradigm.

3. Application to a linear sum of components

a. Construction of the synthetic dataset

The synthetic dataset used in this study is generated to mimic the results of that are often expected from a PCA decomposition. For this purpose, $Q = 6$ components representing six different dynamical phenomena are used. Each component is described by a different temporal base function $\boldsymbol{g}_i$ (solid lines in Figure 3) constructed from composites of sinusoids with different frequencies and phases. These base functions have been normalized to give a standard deviation of unity; i.e. each component accounts for a similar amount of temporal variance. The temporal dimension of these base functions is taken to be $N = 365$ (e.g., one year of daily data). A spatial resolution of $2.5^0 \times 2.5^0$ is chosen, corresponding to $M = 144 \times 72 = 10368$ pixels. Finally, the dataset, $\boldsymbol{X}_j{}^t = \{x(j) \in R^N \ ; \ \ j = 1, \ldots, M\}$, where R^N is the space of real vectors of dimension N, $N = 365$ and $M = 10368$, is formed from the time series $x(j)$ for each pixel j by the linear sum of the base functions, $\boldsymbol{x}(j) = \boldsymbol{g}_1\sigma_1(j) + \ldots + \boldsymbol{g}_Q\sigma_Q(j) + \varepsilon$ (linear model of Eq. 1). The term ε is a Gaussian-distributed noise (zero mean and standard-deviation of 0.5), representing very noisy data.

The $\{\sigma_i(j) \ ; \ \ i = 1, \ldots, Q\}$ indicate the strength of each component, i, at each pixel, j. These strengths are constructed to have a geographical Gaussian distribution, giving a different ellipsoidal distribution for each component (left column in Figure 4). Land contours are shown for easier description of the modes. One of the components has two peaks in its spatial distribution to represent a teleconnection pattern (map of component 1 in left column of Figure 4), so the total number of ellipsoidal peaks is seven. Also, the geographical extent of two of the components overlaps in the Indian Ocean (maps of components 4 and 6 in left column of Figure 4) to complicate the component extraction process.

The variance explained by the $Q = 6$ components and the variance from the added noise are shown in Table 1. The variance explained by the components represents 67 % of the total variance and the noise represents 33 %. The variance of a component results from the combination of the temporal variability of the base function as a function of amplitude (that has been normalized here) and frequency, and the spatial extent of the component.

b. Results of PCA and ICA

The PCA components are determined by computing the matrix $\boldsymbol{J}_0$. The best number of PCA components to extract is determined by observing the spectrum of cumulative percent of variance explained by the PCA components (Figure 5). The first $Q = 6$ PCA components represent 67.7 % of the total variance and the 359 remaining components that explain 32 3 % of the total variance represent the noise in the dataset. The temporal PCA base functions (crossed lines in Figure 3) are each compared with the real base function to which it best corresponds. PCA base functions 2, 3 and 4 provide a relatively good estimate of the true functions, although there are some errors near the peak values. PCA temporal base functions 1, 5 and 6 (low frequencies) are much worse fits. In particular, higher frequencies have been mixed in with the real base functions. The corresponding PCA score maps are defined at pixel j by the values $(\sigma_1, \ldots \sigma_Q)(j) = \boldsymbol{J}_0 \cdot \boldsymbol{X}_j{}^\star$ and are shown in Figure 4 (middle column), where $\boldsymbol{X}_j{}^\star$ is the j^{th} column of data matrix $\boldsymbol{X}$. We see that the PCA (or EOF) technique confuses elements from the different components, the general mixing problem, such that all of its components exhibit many more geographic peaks than in the real components. Even if the corresponding PCA temporal base function is relatively well retrieved, the corresponding score map still exhibits the mixing problem (see PCA base function No 2 in Figure 3 and the corresponding PCA score map in Figure 4). One cause of the mixing is well illustrated in Table 1 where the variance explained by each PCA components is compared to the variance of the actual components. The first PCA component explains 24.4 % of the total variance, which is much more than the true variance of 13.3 %. The variance maximization constraint on the solution in PCA means that the first component is the mixture of many true component variabilities, whereas the 6^{th} PCA component represents only 3.1 % which is a considerable under-estimate of the real value of 10 %. The noise level estimate of 32.3 % is a good estimate, but its small under-estimate of the real noise is due to the projection of noise into the first 6 PCA components (representing 0.7 %).

This mixing tendency of the PCA could suggest many more teleconnections in observations than are actually present. Since all six components contribute the same amount of variance, the PCA technique has combined many of the actually-separate components into several of its components, trying to maximize the amount of variance explained by each. However, the method is then compelled to alternate positive and negative values to compensate for having too much variance when the components are added back together. This effect is especially apparent for the overlapping components in the Indian Ocean: two PCA base functions possess broad central peaks spanning the geographic distribution of both of the real components and two others possess, in this same location, two opposite-signed peaks (see PCA score maps of components 1,4, 5 and 6 in Figure 4, middle column). A similar projection of real components into more than one PCA component occurs when a geographically isolated mode moves during the time period (Kim and Wu 1999). Moreover, the component with two peaks in the Americas, representing a real teleconnection, shows up in four of the PCA components (components 1, 3, 4 and 6 in Figure 4, middle column), but mixed with other components as well, suggesting teleconnections between the Americas and the South Atlantic and Indian Oceans that do not exist.

For the ICA, a "whitening" procedure is applied first using the PCA solution: the observed data, $\boldsymbol{x}(j)$, are projected onto the first $Q = 6$ PCA components using the matrix $\boldsymbol{J}_0$. The ICA technique is then applied to the "whitened" data, $\tilde{\boldsymbol{x}}(j) = \boldsymbol{J}_0 \cdot \boldsymbol{x}(j)$ (dimension $Q = 6$ instead of $N = 365$). This is equivalent to performing an oblique rotation on the PCA initial solution, see (Nadal et al. 1999, Aires et al. 2000). Thus the 6 ICA extracted components explain the same amount of variance as the 6 PCA components (67.7 %).

The six ICA base functions are shown in Figure 3 (dotted lines). The ICA base functions are very similar to the real base functions; this comparison shows how the ICA technique has corrected its first guess (the PCA solution) to be closer to the true temporal base functions. The additional information conveyed by the requirement of statistical independence is nicely illustrated: the ICA solution is better than the PCA solution for all six components. The ICA score maps are presented in Figure 4 (right column). The presence of large-amplitude noise amplifies the ambiguity, producing weak mixing of the components (see the score map for component 5, for example); but generally, the components are well retrieved and separated, even the teleconnection mode (ICA component 1 in Figure 4, right column) and the two overlapping modes in the Indian Ocean (ICA component 4 and 6 in Figure 5, right column). The ICA score maps are always an improvement over the PCA solution. When the noise is removed, the ICA separates the original six modes very cleanly with very little mixing; thus, in practice, if the noise amplitude is significantly smaller than the signal amplitude, the ICA solution is very close to the real solution.

Table 1 shows that the variance explained by the ICA components is closer to the real solution than the PCA components. Differences between the true and ICA explained variance for each components are less than 0.6 %. Discrepancies are the result of the projection of some part of the noise into the ICA components.

4. Concluding remarks

PCA (or EOF analysis) provides an economical way to summarize or characterize the complex time and regional variations of a geophysical parameter over the whole globe with only a few functions. When the total time variance is dominated by one or two separate orthogonal modes that add approximately linearly and with different variances, then this analysis can also correctly identify such modes. However, PCA is now routinely applied to the study of atmospheric observation anomalies, which are defined as deviations from the average leading mode, the annual cycle. There are three problems. (1) The annual cycle, itself, cannot always be described by a single EOF mode (e.g., Rossow et al. 1993), indicating that its phase and/or regional distribution are varying over the time record so that the **mean** annual cycle is only an approximation to the actual annual cycle in a given year. (2) The several anomalies that have been identified by removal of the mean annual cycle generally only account for a few percent of the total variance and are not a lot larger than data noise. (3) The fluid dynamics of the atmospheric circulation strongly couples motions on different space and time scales, i.e., the separate modes combine **non-linearly** to produce the total time signal. The latter fact means that the annual cycle is constantly interacting non-linearly with atmospheric variations on a whole range of other time scales, so that the removal of the mean annual cycle from a dataset leaves behind a residual that is only a partial representation of these physical interactions, producing an artificial mode from the analysis method. Thus, there is every reason to suspect that the assumptions that underlie the PCA are violated when applied to atmospheric circulation anomalies. This leads to the worry that the anomalies that have been identified and studied are not physical, but are either created by the mixing of many separate physical modes having similar amplitudes, but different regional distributions and interacting non-linearly, or by the mixing (aliasing) of the annual cycle into other modes by observation errors. Our simple example shows how extra modes can be generated by regional overlaps between two different modes and how spurious teleconnections can be found where none exist or distorted where they do exist, even when the modes add linearly, (i.e., favorable condition for the PCA).

Our simple linear example also shows the potential of the ICA technique for separating a complex signal in a more meaningful way. The mixing problem inherent in the PCA technique and the artifacts prod ıced by the orthogonality and maximum variance constraints of PCA are avoided with the ICA approach. Moreover, the use of higher-order statistics to determine statistical independence assumes much less about the character of the parameter distributions than the PCA. In the case where the components of the system are linked in a certain way (the climate is certainly closer to this model), the ICA would be able to extract components that are a kind of prototype, defined to be as statistically distinct as possible, for an optimal description of the variability in the observations. This means that ICA, solving the mixing problem, is more suitable for global studies, which is not the case of PCA.

As with the classical PCA technique, this first ICA algorithm is not able to deal correctly with propagating components. But the ICA paradigm may be a sufficiently general concept to be used in a more sophisticated way like complex ICA or nonlinear ICA. Our experiment on synthetic data, where the solution of the component extraction problem is known, encourages further work to extend the ICA paradigm to non-linear cases: this requires development of non-linear solution algorithms and testing for cases where the combination of modes is non-linear, when components are physically linked, and for cases with propagating modes.

Appendix: Principal steps of the algorithm

We adopt here the linear model $\boldsymbol{x} = \boldsymbol{G} \cdot \boldsymbol{\sigma}$, where $\boldsymbol{x}$ is the observation, $\boldsymbol{G}$ is the base function matrix and $\boldsymbol{\sigma}$ is the vector of components to estimate. The goal of the statistical decomposition technique is to estimate a matrix $\boldsymbol{J} = \boldsymbol{G}^{-1}$, the filter matrix, using only a dataset of observations $\{\boldsymbol{x}^e \; ; \; e = 1, \ldots, E\}$, where E is the number of samples in the dataset. With this matrix $\boldsymbol{J}$, for each observation $\boldsymbol{x}$, the components $\boldsymbol{\sigma}$ are estimated by $\boldsymbol{h} = \boldsymbol{J} \cdot \boldsymbol{x}$, and the base function matrix $\boldsymbol{G}$ is estimated by the inverse matrix $\boldsymbol{J}^{-1}$.

The principal steps of the time series analysis by the ICA technique are:

- **Pre-processing of dataset**: The dataset $\boldsymbol{X}_j{}^t = \{\boldsymbol{x}(j) \in R^N \; ; \; j = 1, \ldots, M\}$, where t is the time index and j is the space index (geographical locations), often requires a pre-processing step:

  - spatial, temporal or spatio-temporal interpolation to resolve missing data problems,
  - filtering of data to suppress undesirable frequencies,
  - de-trending to obtain a stationary data,
  - removing the annual cycle to examine interannual anomalies.

None of these steps is required. For example, we have commented in the paper on the dangers of removing the annual cycle.

- **Chose the space for the decomposition**:

- in time, which is the approach we have adopted in our study: $\boldsymbol{x}(j) = \boldsymbol{g}_1\sigma_1(j) + \ldots + \boldsymbol{g}_Q\sigma_Q(j) + \boldsymbol{\varepsilon}$,
- in space,
- in frequency,
- in a mixture of these spaces.

The observations (a time series or a geographical field, ...) are noted, in the following, by the d-dimensional vector $\boldsymbol{x}^e$ and the dataset is $\{\boldsymbol{x}^e \ ; \ \ e = 1, \ldots, E\}$.

• **Center the dataset**: The observation mean $< \boldsymbol{x}^e >$ is removed from the dataset: $\boldsymbol{x}^e \leftarrow \boldsymbol{x}^e - < \boldsymbol{x}^e >$. This step is necessary for statistical techniques where data are supposed to have zero-mean.

• **Normalize the dataset**: If the user wants to put the same statistical weight on each coordinate of the observation $\boldsymbol{x}^e$ (that could be a date in time decomposition, a pixel location in space decomposition, ...): observations in the dataset are normalized by the standard-deviation vector $\boldsymbol{x}^e \leftarrow \boldsymbol{x}^e / \boldsymbol{e}_x$.

• **Eigen-vector decomposition**: The covariance (or correlation, in the case of normalized observations) matrix $< \boldsymbol{x}^t \cdot \boldsymbol{x} >$ is estimated from the dataset. The eigen-values $\boldsymbol{\Lambda}$ (diagonal matrix) and the eigen-vector matrix $\boldsymbol{V}$ of $< \boldsymbol{x}^t \cdot \boldsymbol{x} >$ are then computed using a classical numerical routine. The number of PCA or ICA extracted components Q is chosen by observing the spectrum of eigen-values.

• **PCA solution**:

- The $d \times Q$ PCA base function matrix $\boldsymbol{G}_{PCA}$ contains in its columns the first Q eigen-vectors of $\boldsymbol{V}$ (the columns of $\boldsymbol{V}$ represent time series in the time decomposition, and geographical field in the space decomposition, ...).

- Since, by definition, $\boldsymbol{V}^{-1} = \boldsymbol{V}^t$, the filter PCA matrix $\boldsymbol{J}_{PCA}$ is equal to the transposed $Q \times d$ base function matrix $\boldsymbol{G}_{PCA}$. Then, the extracted components $\boldsymbol{h}$ that estimate the true components $\boldsymbol{\sigma}$ are the projection of the observation $\boldsymbol{x}$ onto the filters: $\boldsymbol{h} = \boldsymbol{J}_{PCA} \cdot \boldsymbol{x}$.

- The first Q eigen-values in $\boldsymbol{\Lambda}$ represents the variability explained by each of the Q components.

- **ICA solution**:

1 Pre-whitening of dataset: The PCA solution is used for a pre-processing data step: the observations $\boldsymbol{x}^e$ are projected into the PCA filters: $\boldsymbol{x}^e \leftarrow \boldsymbol{J}_{PCA} \cdot \boldsymbol{x}^e$. The ICA algorithm is then applied into these Q-dimensional data.

2 The ICA solution $\boldsymbol{J}_{ICA}$ is initialized as the identity matrix $\boldsymbol{I}_{Q \times Q}$. This, associated to the previous whitening step of data, is equivalent to taking the PCA solution as first guess for ICA.

3 For the minimization of the criterion specifying the statistical independence, a stochastical gradient descent algorithm is used. The stochastical principle uses the gradient descent formula iteratively in unique random samples of the dataset, contrary to the classical approach where the gradient descent formula is applied

iteratively to the global dataset (i.e. a deterministic algorithm). The stochastical character of an optimization algorithm allows theoretically, under some constraint not discussed here, for the optimization technique to reach the global minimum of the criterion instead of the local minimum where a deterministic algorithm could be trapped.

4 An observation $\boldsymbol{x}^e$ is randomly chosen in the dataset. The propagation through the neural network (chosen model for the extraction component) is given by: $\boldsymbol{y} = f(\boldsymbol{h}) = f(\boldsymbol{J}_{ICA} \cdot \boldsymbol{x}^e)$, where $f(a) = 1/1 + exp(-\beta \cdot a)$ is the logistic function (β is a parameter controlling the slope of the logistic function, we take $\beta = 2.0$). The FORTRAN routine of this process is:

```
c - - propagation into the neural network
    do $i = 1, d$
        $h(i) = 0.d0$
        do $k = 1, d$
            $h(i) = h(i) + J_{ICA}(i,k) * x^e(k)$
        enddo
        $h(i) = h(i) + bia(i)$
        $y(i) = 1.d0/(1.d0 + dexp(-\beta * h(i)))$
    enddo
```

where *bia* is the classical bias vector in a MLP neural network (not shown in the text for simplicity). We use, in this routine, double precision variables to avoid numerical instabilities.

5 The learning process is then defined as:

```
c - - transitory quantities

    do $j = 1, d$

        $hhh(j) = 0.d0$

        do $k = 1, d$

            $hhh(j) = hhh(j) + J_{ICA}(k, j) * h(k)$

        enddo

    enddo

c - - modification of weights

    do $i = 1, d$

        do $j = 1, d$

            $J_{ICA}(i, j) = J_{ICA}(i, j) + param*$

    &       $(J_{ICA}(i, j) + \beta * (1.d0 - 2.d0 * y(i)) * hhh(j))$

        enddo

        $bia(i) = bia(i) + \beta * (1.d0 - 2.d0 * y(i))$

    enddo
```

where *param* is the learning parameter of the gradient descent optimization (we take $param = 0.0005$).

6 Stopping criterion: many criteria can be used to define when stopping the previous learning steps. The simplest criterion is to determine a priori the number of learning steps. A better adequate criterion is the measure of the difference between solution $\boldsymbol{J}_{ICA}$ at time t and at time $t+1$: if this difference is low, the algorithm has converged. Another stopping criterion is the measure of the statistical independence of the extracted components $\boldsymbol{h}$: cumulants (i.e. additive higher-order moments) are a practical way to do that, but this approach is still computationally expensive. The learning algorithm returns to step 4 until the stopping criterion is reached.

- **Analysis of results**: When the matrix $\boldsymbol{J}_{ICA}$ has been determined by ICA, the global ICA filters (taking into account the PCA pre-processing) are defined by the $Q \times d$ matrix: $\boldsymbol{J}_{GLO} = \boldsymbol{J}_{ICA} \cdot \boldsymbol{J}_{PCA}$

  - The projection of data is used to estimate the components: $\boldsymbol{h} = \boldsymbol{J}_{GLO} \cdot \boldsymbol{x}^e$
  - The $d \times Q$ ICA base function matrix $\boldsymbol{G}_{GLO} = \boldsymbol{J}_{GLO}{}^{-1} = \boldsymbol{G}_{PCA} \cdot \boldsymbol{J}_{ICA}{}^{-1}$ is normalized to obtain normalized ICA base functions, as in PCA approach.
  - Computation of explained variance of each of the base functions.

Acknowledgments. We are grateful to Dr. David Rind and Dr. Ronald L. Miller for their helpful comments. This work was supported by special funding provided by Dr. Robert

J. Curran, NASA Climate and Radiation Program.

References

Aires F., A. Chédin, and J.-P. Nadal, 2000: Independent Component Analysis of Multivariate Times Series: Application to the tropical SST variability, *Journal of Geophysical Research*, accepted for publication.

Bell A. J., and T. J. Sejnowski, 1995: An information-maximization approach to blind separation and blind deconvolution, *Neural Computation*, **7**, 6, 1004–1034.

Burgers, G., and D.B. Stephenson, 1999: The "Normality" of El Nirõ, *Geophysical Research Letters*, **26**, 8, 1027–1030.

Horel, J., 1981: A rotated principal component analysis of the inter; nnual variability of the northern hemisphere 500 mb height field, *Monthly Weather Review*, **109**, 2080–2092.

Kim, K.-Y., and Q. Wu, 1999: A comparison study of EOF techniques: analysis of nonstationary data with periodic statistics, *Journal of Climate*, **12**, 185–199.

Lorenz, E., 1951: Seasonal and irregular variations of the northern hemisphere sea-level pressure profile, *Journal of Meteorology*, **8**, 52–59.

Nadal J.-P., and N. Parga, 1994: Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer, *Computation in Neural Systems*, **5**, 565–581.

Nadal J.-P., and Parga N., 1997: Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches, *Neural Computation*, **9**, 7, 1421–1456.

Nadal J.-P., E. Korutcheva, and F. Aires, 1999: Blind source separation in the presence of weak sources, *Neural Networks*, submitted.

North, G.R., 1984: Empirical Orthogonal Functions and Normal Modes, *Journal or the Atmospheric Sciences*, **41**, 5, 879–887.

Richman, M., 1981: Obliquely rotated principal components: an improved meteorological map typing technique ?, *Journal of Applied Meteorology*, **20**, 1145–1159.

Richman, M., 1986: Rotation of principal components, *Journal of Climatology*, **6**, 293–335.

Rossow, W.B., A.W. Walker, and L.C. Garder, 1993: Comparison of ISCCP and other cloud amounts, *Journal of Climate*, **6**, 2394–2418.

Filipe Aires and Willian B. Rossow, NASA Goddard Institute for Space Studies, 2880 Broadway, New-York, NY 10025, USA. (e-mail: faires@giss.nasa.gov; wrossow@giss.nasa.gov)

Alain Chédin, CNRS Laboratoire de Météorologie Dynamique, École Polytechnique, 91128 Palaiseau Cedex, France. (e-mail: chedin@jungle.polytechnique.fr)

Received ...., 2000; revised ...., 2000; accepted ...., 2000.

To appear in the *Journal of Climate*, 2000.

Figure 1. Problems encountered by PCA when observations have dimension 2 (coordinates X and Y) and come from two components defining ellipses E1 and E2, the line D represents the first PCA axe defining the first PCA component: A) mixing due to the maximum explained constraint, B) indeterminacy when two components have same variance, and C) mixing due to the non-orthogonality of components.

Figure 2. T ıe component extraction model: the perceptron architecture, where $\boldsymbol{x}$ is the observation, $\boldsymbol{h}$ is the extracted component vector and $\boldsymbol{h}$ is the ouput network.

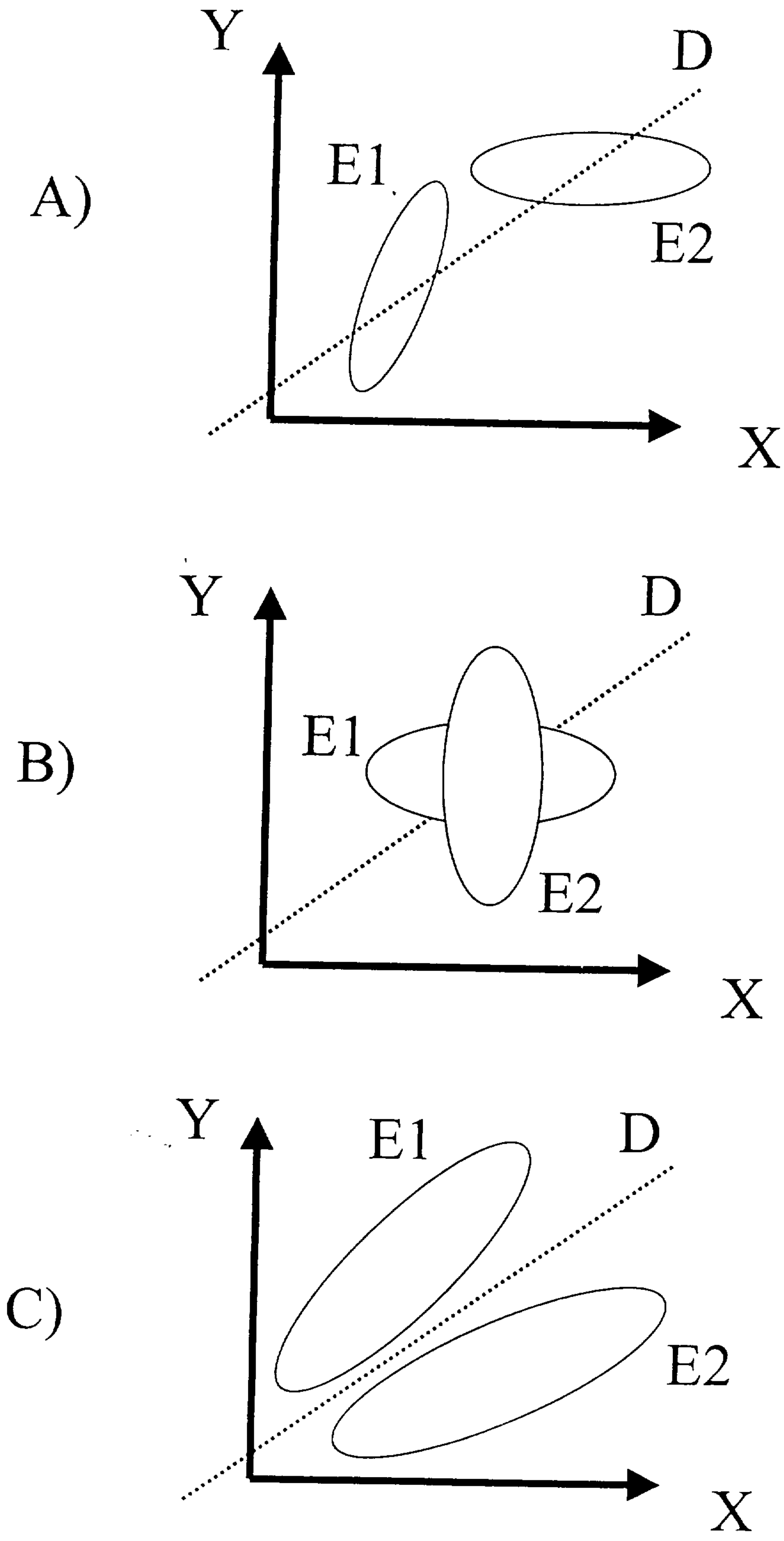
Figure 3. Temporal base functions g_i: ACTUAL (solid lines), PCA estimates (crossed lines) and ICA estimates (points).

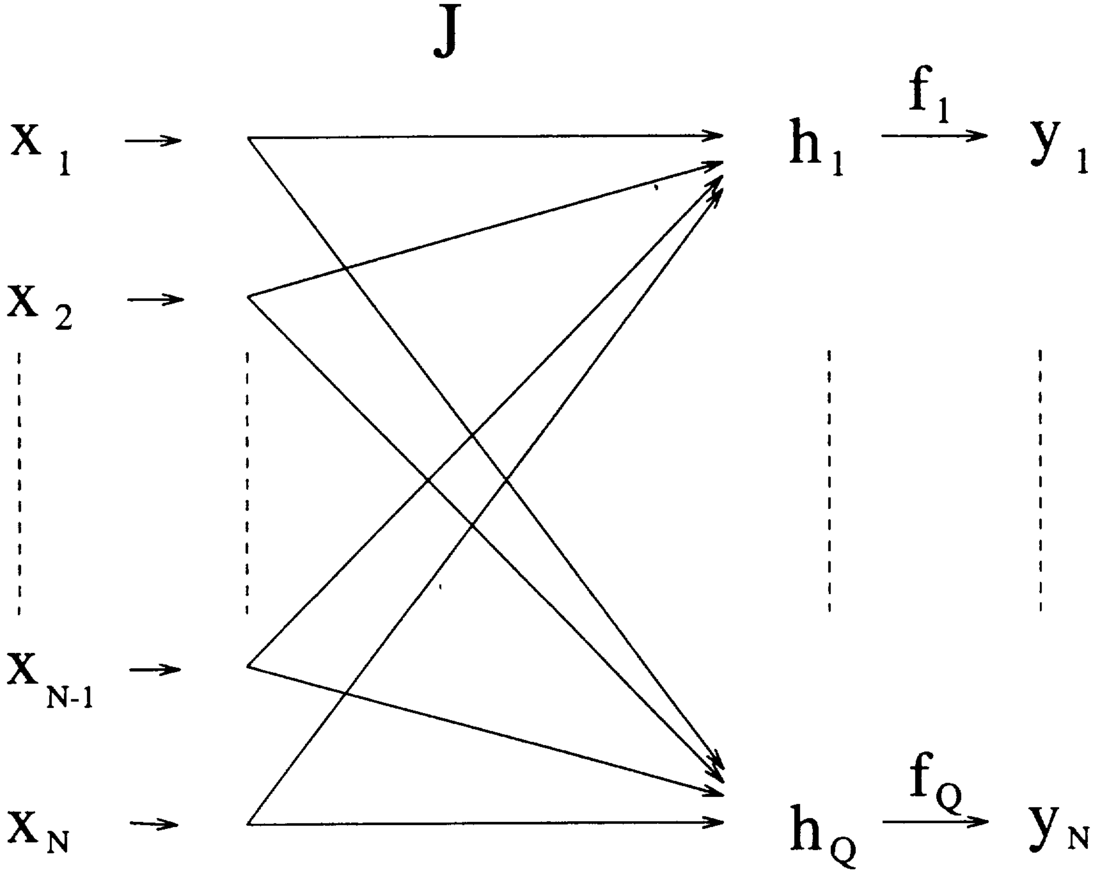
Figure 4. The components score maps of actual components σ_i (left column), of the PCA extracted components h_i (middle column), and of the ICA extracted components h_i (right column): components number 1-6 from the top to the bottom, colors from the blue to the red (rainbow) representing range 0-1 for actual amplitudes (left colum) and -3 to +3 for extracted components amplitudes (middle and right column)

Figure 5. Cumulative percent of explained variance by the PCA components

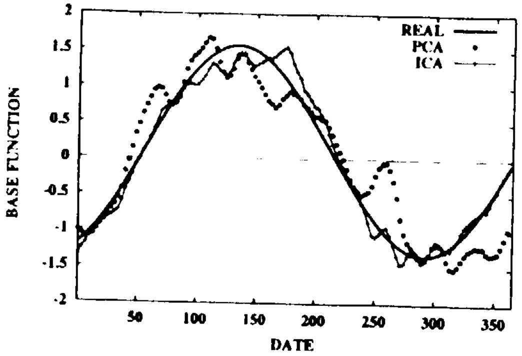
Table 1. Variance explained by Noise, REAL, PCA and ICA components

Component	REAL	PCA	ICA
1	13.3	24.4	12.7
2	12.6	14.5	13.0
3	10.7	10.7	11.3
4	10.7	8.8	10.2
5	10.7	6.2	10.3
6	10.0	3.1	10.0
Noise	33.0	32.3	32.3

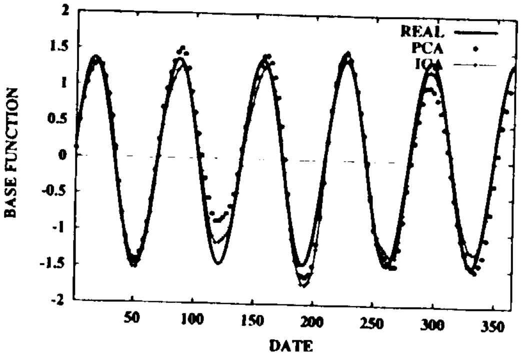

A)
Y
D
E1
E2
X
B)
Y
D
E1
E2
X
C)
Y
E1
D
E2
X


J
x 1
x 2
x N-1
x N
h 1
h Q
f 1
f Q
y 1
y N

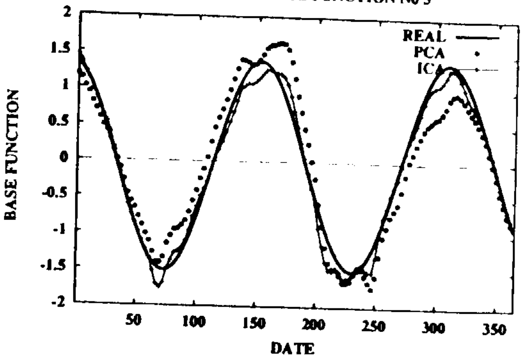
TEMPORAL BASE FUNCTION No 1



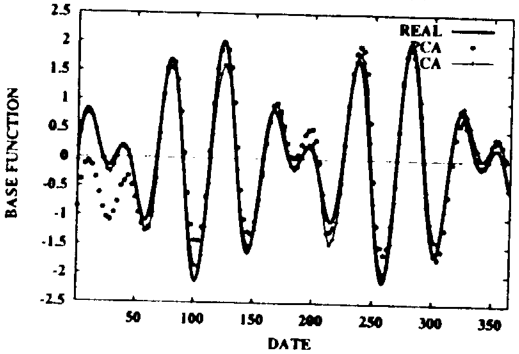
TEMPORAL BASE FUNCTION No 2



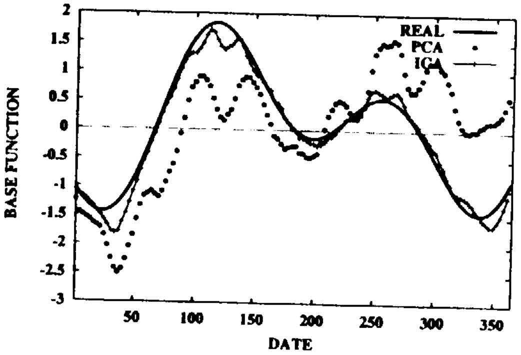
TEMPORAL BASE FUNCTION No 3



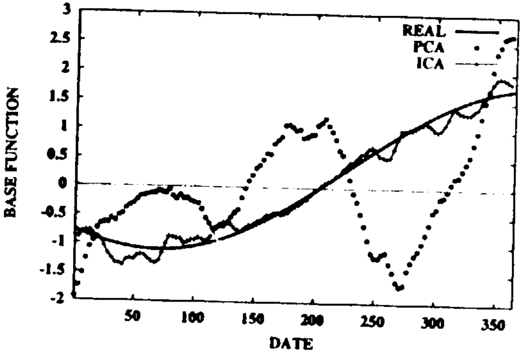
TEMPORAL BASE FUNCTION No 4

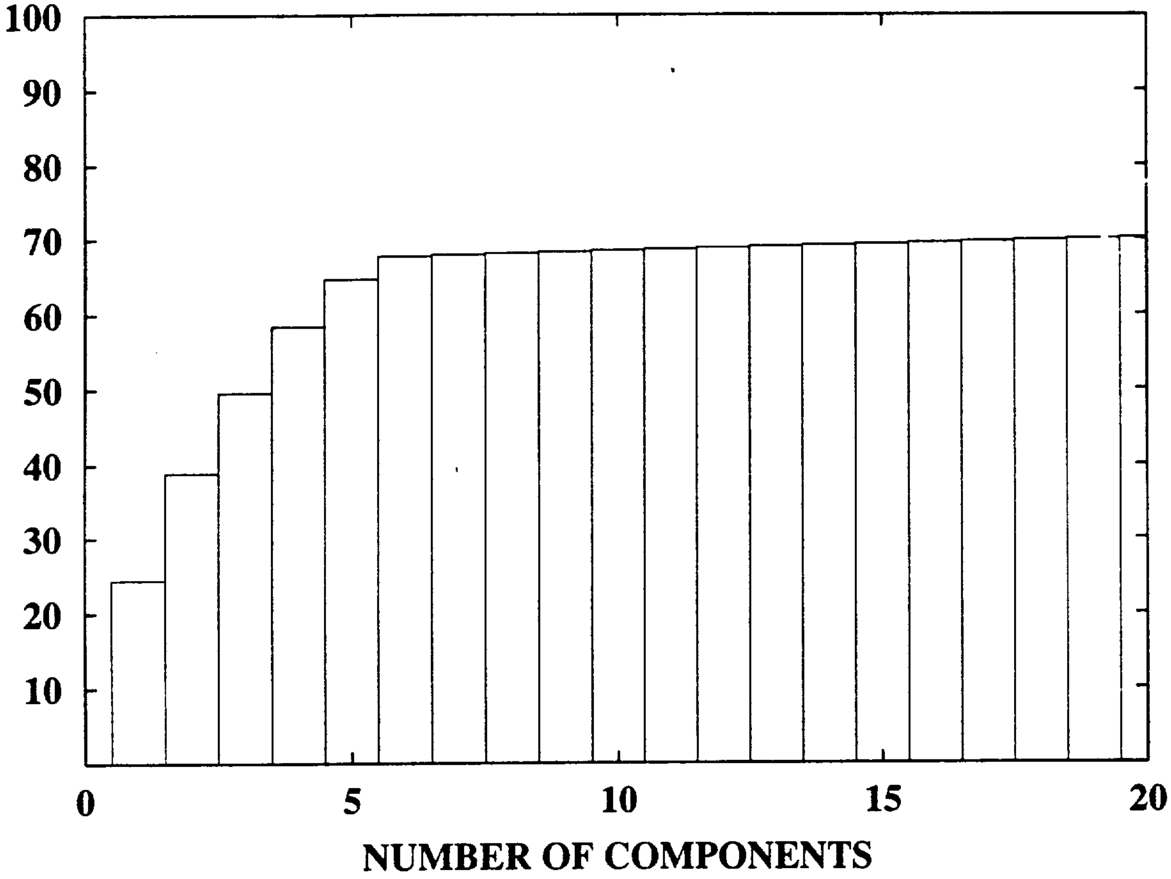


TEMPORAL BASE FUNCTION No 5



TEMPORAL BASE FUNCTION No 6





100
90
80
70
60
50
40
30
20
10
0
5
10
15
20
NUMBER OF COMPONENTS